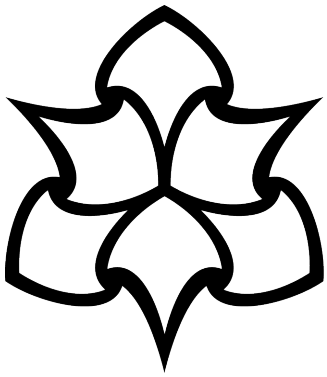

Understanding Insider Threat Attacks using Natural Language Processing: Automatically mapping organic narrative reports to existing insider threat frameworks

Katie Paxton-Fear[1][0000−0002−4472−8955], Duncan Hodges[1][0000−0002−0660−8776], and Oliver Buckley[2][0000−−0003−1502−5721]

Centre for Electronic Warfare, Information and Cyber, Cranfield University, Defence Academy of the United Kingdom, SN6 8LA
{k.paxton-fear,d.hodges}@cranfield.ac.uk
School of Computing Sciences, University of East Anglia, Norwich NR4 7TJ
o.buckley@uea.ac.uk

Abstract. Traditionally cyber security has focused on defending against external threats, over the last decade we have seen an increasing awareness of the threat posed by internal actors. Current approaches to reducing this risk have been based upon technical controls, psychologically understanding the insider's decision-making processes or sociological approaches ensuring constructive workplace behaviour. However, it is clear that these controls are not enough to mitigate this threat with a 2019 report suggesting that 34% of breaches involved internal actors. There are a number of Insider threat frameworks that bridge the gap between these views, creating a holistic view of insider threat. These models can be difficult to contextualise within an organisation and hence developing actionable insight is challenging. An important task in understanding an insider attack is to gather a 360-degree understanding of the incident across multiple business areas: e.g. co-workers, HR, IT, etc. can be key to understanding the attack. We propose a new approach to gathering organic narratives of an insider threat incident that then uses a computational approach to map these narratives to an existing insider threat framework. Leveraging Natural Language Processing (NLP) we exploit a large collection of insider threat reporting to create an understanding of insider threat. This understanding is then applied to a set of reports of a single attack to generate a computational representation of the attack. This representation is then successfully mapped to an existing, manual insider threat framework.

Keywords: Insider Threat · Natural Language Processing · Organic Narratives.

1 Introduction

An insider threat can be defined as 'a current or former employee, contractor, or business partner who has or had authorised access to an organisation's network,

system, or data and intentionally exceeded or misused that access in a manner that negatively affected the confidentiality, integrity, or availability of the organisation's information or information systems' [6] and represents a major security risk to organisations. A 2019 report compiled by Verizon suggests that 34% of all data breaches involved an internal actor and 15% of actions taken during a breach were misuse of a system by authorised users [32]. Insiders were shown to be particularly common in specific sectors such as the Public Sector (30% of all breaches), Finance (36% of all breaches) and Healthcare (59% of all breaches). It is clear that despite the increased availability of technical solutions and the increased awareness of the insider threat, insiders are still able to plan and commit attacks.

The current state of the art approaches to understand, and ultimately prevent insider threat, usually consider either technical [11,1,8], psychological or sociological [17,12,4] approaches with a small number of models which encapsulating a number of these factors (e.g. [27]). Within this domain, technical solutions usually aim to restrict or detect insider threat activity, distinguishing malicious activity from non-malicious activity; this is often very challenging as, by definition, the insider activity can closely resemble normal, 'everyday' activity. Alternatively, the psychological and sociological approaches aim to understand the factors and decision-making processes involved in initially becoming a threat and then the process of moving from a threat to committing a malicious act.

However, in isolation these approaches are usually not enough to fully understand and contextualise the threat from an insider attack within an organisation. There are also a number of frameworks which attempt to bring together these approaches to create a more holistic understanding of the problem, acknowledging that the threat from insiders is a nuanced socio-technical problem [7,19]. These frameworks can provide an abstract appreciation for the relevant factors but it can be challenging to map this understanding to a tangible set of mitigations which can reduce the risk from insider threat.

Following a security breach from an insider it is important to gather information regarding the incident, gathering data from co-workers including the individual's peers, juniors and seniors, human resources and those in a staff management role, in addition to IT or security personnel. This represents a diverse community of individuals all of whom may have important information pertaining to the incident. With research showing that these individuals are willing to write reports about an incident when one has occurred, giving investigators important information regarding an attack [10]. This research acknowledges that casting this evidence base into an insider threat framework will help to evaluate the incident, however the mechanism by which this model integration could lead to erroneous sense-making.

Requiring this variety of individuals to cast their understanding into framework, may bias individuals' recollections of the events and exhibited behaviours as individuals attempt to cast their story into the model framework. In addition there is a cognitive load associated with performing this activity, particularly with those who are not familiar with the framework being used. These factors

can result in a view of the incident which is distorted to fit into existing understandings of insider threat, rather than accurately encapsulating the events associated with the insider attack.

In this paper we describe a computational approach to collecting reports of insider attacks that uses organic narratives describing the insider attack to build a model representation of the incident. It is important to understand that these organic narratives are created by non-experts and represent 'free' text written in 'natural' language. The aim is that this reduces both the security expertise, and the cognitive load required to contribute information to an investigation of an insider threat attack. This computational approach is unsupervised and delivers a model representation of the attack derived from a corpus of organic narratives, this model is free from assumptions about how insider threat attacks have traditionally been committed allowing a rich understanding of new and emerging attacks.

In this paper we first discuss the background to this work, then introduce our approach and describe an experiment validating the ability for our approach to produce models that accurately represent insider threat. We finally discuss the implications of our approach and highlight future research directions exploiting this technology.

2 Background

2.1 Insider Threat

As previously discussed insider threat is defined as 'A current or former employee, contractor, or business partner who has or had authorised access to an organisation's network, system, or data and intentionally exceeded or misused that access in a manner that negatively affected the confidentiality, integrity, or availability of the organisation's information or information systems' [6]. These attacks can be more dangerous than attacks from external actors as insider often have access to privileged information, valid credentials and knowledge of potential security systems. Security compromises form insider actions often follow 4 primary archetypes [6]. The first three archetypes are associated with a malicious insider, these include insider fraud, insider IT sabotage and insider IP theft; the final archetype is the so-called unintentional insider threat which can be defined as an insider who *'... through action or inaction without malicious intent unwittingly causes harm or substantially increases the probability of future serious harm'* [13].

The current approaches for understanding insider threat can be characterised into three general approaches; the first is a technical approach that aims to understand the technical artefacts an insider leaves on an IT system and from this identify or block activity associate with an insider threat, (e.g. [25]). The challenges in this approach is that by the nature of insider activity, it often closely resembles everyday activity and is often perpetrated by those who have the knowledge to reduce their exposure to technical detection mechanisms.

The psychological and sociological approaches aim to understand any predisposition resulting from personality which increases the risk of becoming an insider threat (for example, [22]), additionally, there are other approaches which attempt to model the decision-making during the transition of an employee to becoming an insider threat (for example, [14]).

The socio-technical characterisation of the insider threat is typically brought together in the third approach to understanding insider threat, using a framework or model in order to understand the complex and nuanced interactions between the different elements. The relationships between these elements associated with insider threat can form complex feedback loops, which attempt to model the individual, the environment or organisation and the security incident itself. These typically acknowledge that technical controls are often insufficient to manage the risk from insider threat [9] and a holistic understanding of the risk and hence controls is essential to begin to manage the risk.

These models of insider threat (for example [27,20,5,23]) typically identify a set of themes as related to insider threat and then identify connections between the themes, typically highlighting causal relationships. These relationships demonstrate opportunities for mitigations or detection, whilst also highlighting possible feedback loops that look to increase the likelihood of, for example a successful attack or an employee becoming a threat in the first place. In an academic sense these models are very useful for understanding the interactions between a variety of environmental and technical elements as well as the effect of individual differences, however when translated into an organisational workplace or as a tool for post-hoc exploration of an incident they typically require a knowledgeable security professional to carefully explore the evidence from a number of sources and then contextualise this evidence within the model.

There are two approaches to post-hoc exploration of an incident using a model, the first involves collecting the evidence surrounding the incident directly in a model with individuals who are providing evidence doing so directly into a model representation. This relies on those reporting the incident to be able to contextualise their observations into a model, insider attacks are often nuanced and complex [7] which can mean this contextualisation is very challenging and requires a significant amount of domain knowledge, it can also result in observations of incident being unconsciously adjusted to fit the model, rather than accurately represent the incident.

The alternative approach to post-hoc exploration of an incident is for an investigator to collect a number of different reports of the incident and then contextualise these within a model. This potentially represents a more accurate approach as by encouraging individuals to report their observations as organic narratives written in 'natural' language allows an individual to represent the incident to the best of their ability. However, this approach requires a significantly skilled individual to gather this information and contextualise it within a model, this will typically be a security domain expert. They will likely have individual bias as to what is expecting to be seen, based on both previous experience and the common threads within the model itself. This confirmation bias

will potentially result in a model representation which is a convolution of both the individual reports and what the security professional is expecting to see and has seen in the past.

The approach outlined in this paper attempts to support this second approach to the post-hoc exploration of a security incident by automatically generating a model representation of the reports of an insider attack. In our approach we have focused on taking organic narrative reports, these are a narrative that links all actions and actors which provides more information about the incident and the protagonists as the narrative continues (rather than an episodic narrative which considers a report constructed of a number of small incidents) — in our experiments performed with non-experts it is clear that non-experts tend to construct reports of these attacks as organic narratives. By computationally summarising these documents into a model representation our approach simplifies this task and also removes elements of cognitive bias from the model synthesis, we also anticipate that it should be able to resolve new attack vectors previously unseen as the approach purely considers the corpus of natural, organic narratives from the attack rather than previous examples of attacks.

2.2 Natural Language Processing

Natural Language Processing (NLP) is a collection of methods to computationally process and understand human language [24]. NLP is used for many applications that involve natural language such as machine translation [16], question answering [16], information retrieval [30], speech recognition [31] and speech production [31] . These allow computational activities to infer, enrich or perform other operations on human-generate texts.

These approaches typically exploit a large corpora (collection of documents) to build statistical, computational models of the text. These models can then either be used to generate new insight from an existing corpus, or by applying the models to previously unseen text and generate new insight as evidence is gathered.

One popular approach to information retrieval and understanding the content large documents is the use of Topic Modelling [18]. A topic model groups statistically related words, in essence creating a statistical representation of what a piece of text is 'about'. The most common algorithm used to create topic models is the LDA (Latent Dirichlet Allocation) algorithm [3]. The LDA algorithm assumes that every document contains a number of topics, and every unique word has a probability of being in every topic. Some words are more discriminative than others, meaning the presence is more indicative of a text being related to a particular topic. It should be noted that this is an unsupervised technique, individual topics are not 'curated' they are statistical representations that elements to emerge from the corpus.

It is common practice in Natural Language Processing to perform a set of preprocessing tasks to normalise and prepare the corpus. This ensures that the topic modelling approach is efficient and also increases the performance of the model itself. First stopwords are removed, for example an English stopword

list contains the most common English words that do not provide additional information for topic modelling but slow down the training process, these include words such as: 'the', 'a', 'to', 'of'. Stopwords can also be domain specific, as some words in certain contexts can similarly provide no additional information for example when analysing news articles potential stopwords could include 'BBC' or other news organisations [21]. Next the tense of text is normalised using stemming, for example this process normalises 'walk', 'walking' and 'walked' to 'walk', for topic modelling tense does not offer additional context [24]. These processes are important to reduce the cost of training models and to ensure the output topic models represent the most important words in the text and are standard practice when using topic models [18].

The use of topic models to identify the relevant topics in human-generated text provides an ideal approach to ingesting reports of insider threat, this allows individuals to generate their own descriptions of what occurred in their own writing style. This approach to creating narratives reduces the cognitive load of those generating the reports, NLP can then be used to extract the topics which appear across the entire set of narratives. We could hypothesise that these topics would include topics related to, for example, the method used to conduct the attack, the potential impact of the attack, information about the individual and even social elements relating to the perpetrator and their interactions with other staff members. This unsupervised approach will create topics which are derived from the statistical relationships between words in the text rather than a security professional who may be influenced by the existing body of knowledge and what is expected to be seen or confirmation bias [26].

3 Method

The method can be separated into a data gathering phase and three main steps. During the data gathering phase 2 corpora are created:

1. A corpus of insider threat cases taken from news articles
2. A corpus of reports relating to a single insider threat case, these are the organic narratives from observers of the incident

Once these corpora have been collected the main three stages of the modelling approach can be applied, first the creation and labelling of a corpus of individual insider threat reports, second the creation and tuning of a final topic model, and finally the creation of the final mapped report corpus. This full method is shown visually in Figure 1.

3.1 Data Gathering

As previously discussed two corpora are required for this method, first a corpus of many different types of insider threat cases, which are used to generate a topic model. This model encapsulates the various elements of an insider attack from different attacks, that could be written about, for example how the

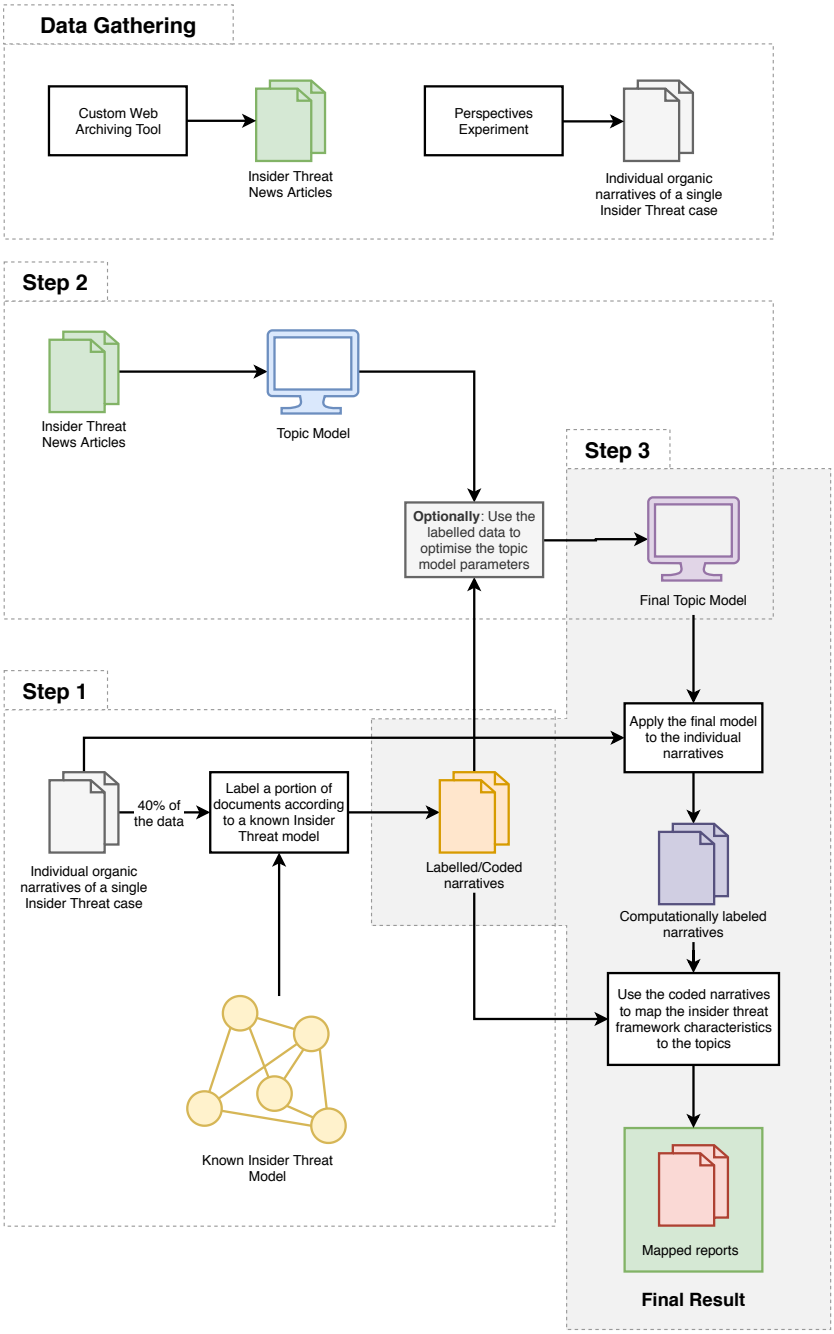

Fig. 1. The full method used to refine organic narratives to a computable model

attack was performed, who perpetrated the attack, whether or not the insider has accomplices etc. These are used to build a model to computationally 'understand' insider threat and is labelled as `Insider Threat News Articles` in Figure 1. The second corpus are the reports of a single incident, these are written as organic narratives and are the corpus we wish to explore.

To create the initial corpus a number of internet sites were identified that report on insider attacks documented on news sites, this includes mainstream news (e.g. BBC News), technology specific sites (e.g. ZDNet) and security specific sites (e.g. Naked Security). A custom web archiving tool was created and these were gathered automatically, this corpus was supplemented by a complex web-scraper that used machine-learning to identify articles about insider threat from news feeds [28]. This creates the final corpus, with a range of insider threat cases all following a similar writing style (that of news articles), with a total document count of 2,700 articles.

The second corpus is the corpus of organic narratives we wish to explore, in application within the workplace these will be gathered from any number of employees witnessing the incidents, the after-effects of the attack or the preceding events, or indeed have simply met the perpetrator. To simulate this corpus of organic narratives a case of insider threat attack was identified from the literature. In our experiment we chose an example case from Nurse et. al. [27] (Case 1), this case was ideal as the insider threat model had already been applied, there were clear witnesses to the insider attack and the case matched the insider threat archetype 'Insider Fraud' [6], a common insider threat attack. This corpus, therefore, is the same insider threat case written with many different writing styles.

Using this as a basis a dramatic recreation was produced presenting the case from three witnesses creating multiple perspectives on the same event. These three witnesses were presented as audio recordings, one witness presented as a news report of the incident, one was presented as relating to a colleague of the perpetrator the final perspective represented the perspective of the event from an IT professional. These three recordings were co-created with professionals in the respective domains to ensure the perspectives were relevant. Study participants were then asked to listen to or read these perspectives and retell the story in their own writing style. Participants were encouraged to write however they preferred using formal or informal language, bullet points or full sentences and as few or as many details as they wanted. Participants were recruited from Mechanical Turk resulting in a final corpus of 107 documents.

3.2 Step 1: Creating Labelled/Coded Reports

The first stage as outlined in Figure 1 is to create a subset of organic narrative reports of the incident. In order to validate the modelling approach this subset is then manually labelled or coded according to a known insider threat model, in our example the model from Nurse et. al. [27] — this also has the added advantage that since the case study is taken from this same paper we have the case study already in the framework as intended by the original authors. In essence

each sentence was allocated a code based on the element of the insider threat model to which it was discussing, e.g. attack step, vulnerability, organisational outcome. This was cross-coded by five independent security researchers and the code with the highest majority was chosen.

This provided a 'human-coded topic model', and provides a guide to which the final computational topic model could be compared.

3.3 Step 2: Creating and Optimising the Computational Models

The second step of the computational process is to use the large corpus of reports to create a model related in insider threat, this is shown as step 2 in Figure 1. This used topic modelling, specifically the LDA algorithm to discover the topics in the documents and create a model for the topics we expect to see in reports of insider threat. It is worth reiterating that this is an unsupervised technique which is solely guided by the statistical relationships between words in the corpus. Exploiting a large corpus of different insider threat news articles which refer to different events, we expect to draw out topics related to a range of insider threat activity, without being reliant on a single case or archetype.

To create the topic models we must choose a value of k, the number of topics. For a large amount of documents this is likely high, although there are methods to automatically compute potential values of k, such as [2].

In this case we do a custom optimisation step which allows us to automatically generate a potential value of k by comparing the characteristics of the computational topic model to the characteristics of the topic model generated by the human-labelling. We make the following assumption, that two sentences that appear within the same code or label are related and therefore we assume that they should appear together in the same topic (or one topic is a subset of the other). Using this desirable characteristic as a metric we can then tune the model hyper-parameters such that this characteristic is maximised. Therefore, the final topic model could have been computed using alternative methods such as [2,15] or with this custom optimisation step. This forms the final output of step 2, from our corpus the final model was generated with a k (number of topics) of 370, and with the removal of stopwords related to both english language and news domain-specific. It is worth noting that all the 370 topics will not be populated when applied to a single case — these represent the putative topics that the unsupervised approach identified as being present in the corpus.

3.4 Step 3: Applying and Mapping the Computational Model

The third and final step in the process as shown in Figure 1 is to apply the model of insider threat to the organic narratives. To do this we exploit a feature of the topic model called a priori probabilities [29]. This takes a segment of text such as a sentence and calculates the probability of that sentence being 'about' each topic, in our case we simply take the highest probability topic and assign the sentence to this topic. This does mean that some sentences can be difficult to place in a single topic, as there can be multiple high scoring topics. Once we

have applied the model all sentences from the organic narratives are associated with a topic.

To evaluate the modelling approach we then need to associate each of these computational topics to the manually coded topics. Since the manually coded topics used a subset of these organic narratives we can identify the topics which contain the same topics, i.e. if a particular sentence appears in topic Y of the human generated topic model and topic X of the machine generated topic model then we can apply the insider threat label (e.g. attack step) from topic Y to topic Y of the machine generated model. This process allows us to add domain context to the unsupervised computational model. It can be noted that this process is not necessarily always accurate, as it relies on a small subset of the data that has been coded in order to appropriately label each topic, however even with this naive approach to labelling each topic we can see that it performs very well.

The output of this final step and indeed the final result of the whole approach is a corpus of organic narratives associated with one insider attack, each sentence of this corpus is mapped to a 'topic' and each of these 'topics' is mapped to the insider threat framework created by Nurse et. al.[27].

4 Results

In this section we will discuss in depth the results of this process and the performance of model overall. In this we highlight several results from the model, these results are provided as the sentences from the organic narratives which are clustered to one particular topic and the insider threat framework entity to which it is related. A representative subset of these topics is shown in Figures 2 to 7. It is worth noting in this section that the original topic model was only trained on a corpus of insider threat news reports and not from the corpus of organic narratives, hence these results demonstrates the training of a generic model that 'understands' insider threat and it's application to a specific 'instance' of insider threat.

When the topics are evaluated, there are some topics that can link to several characteristics from the original framework or the approach is unable to map the topic to a particular characteristics. There are many reasons this could happen, a sentence can map to multiple characteristics or there is no strong link to an existing characteristic. For example a sentence may contain information regarding the behaviour of an individual, this could be considered historical behaviour or observed physical behaviour, and often a single element will have aspects of both characteristics, an example is shown in Figure 3 where the topic contains both a characteristic of an attack and the vulnerability that is exploited. In Figure 2, there is no strong link between the sentences and a characteristic so it remains unlabelled.

An alternative

In addition, the topic model, since it is unsupervised tends to be more specific than a human, for example Topic 265 in Figure 4 and Topic 146 in Figure 5 show two different topics which contain sentences humans both coded as 'Personality

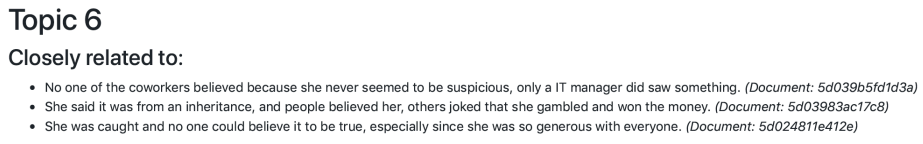

Fig. 2. Topic 6

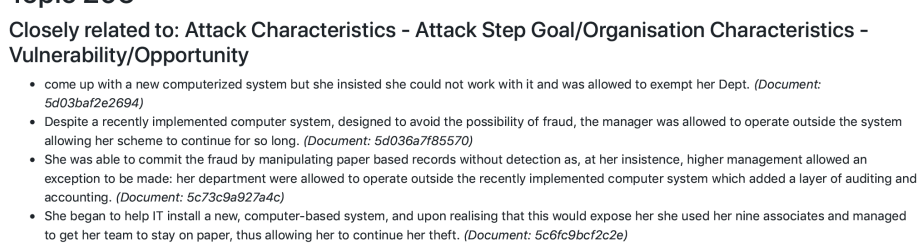

Fig. 3. Topic 205

Characteristics'. However the computational approach has separated these into one topic that described her kindness and influence with a focus on how her colleagues perceived her, the other topic supplements this with additional elements that describe her as 'quirky' and 'flaky'.

These examples demonstrate that the approach is able to correctly identify the various elements and themes associated with insider threat frameworks, however we are also interested in the relationships between these themes. These causal and temporal relationships are particularly important in helping identify opportunities for the reduction of the risk from insider threats as well as better-understanding the 'catalysts' and pathways that enabled a particular incident. A naive approach to linking these themes builds on the identification of the reports as organic narratives in which the report is fundamentally structured in a temporally monotonic manner, with a causal relationship that is linked to this temporal evolution.

From this observation we can then assume that if a sentence that occurs in Topic N is followed by a sentence in Topic Q, there is the potential of a relationship between Topics N and Q. The directed graph of these relationships is shown in Figure 6. Here we can see strong links between the Attack Step and the Asset that was being attacked and the Skill Set of the attacker; between the Precipitating Event and the Attack; between Attack step goal and Attack step and between all Actor Characteristics which are well connected in the graph.

This initial work demonstrates that not only can the approach outlined in this paper be used to in an unsupervised manner map organic narrative reports to an insider threat framework but also begin to identify some of the underlying structure in the framework. This naive approach makes an assumption that there is a temporal and causal link between sentences, whilst this is at times

Topic 265

Closely related to: Actor Characteristics - Personality characteristics

- The employee, a middle level manager, was said to be kind and generous. *(Document: 5d07afae00e24)*
- The news came as a surprise to many who say that she was kind and generous. *(Document: 5d07ad29c4338)*
- The manager is described as kind and generous by her colleagues. *(Document: 5d0481de7bb2e)*
- The news was a shock to the office because the manager had always behaved in a kind and generous way with everyone. *(Document: 5d037c3adef6c)*
- The news came as a shock to the office who had found the manager to be a kind and generous co-worker who frequently supported those in need. *(Document: 5d036a7f85570)*
- Reports say that the manager was a kind and friendly person and no one suspected her of being a thief, despite her oddities. *(Document: 5d02dc8e97cb6)*
- The woman spun a web of lies to hide where the money was coming from and was kind to her friends and staff, enabling her to keep her secret. *(Document: 5c7f9270c1494)*
- But she was kind and generous to her colleagues and was known to frequently support those in need. *(Document: 5c73c9a927a4c)*
- Therefore due to her kindness and generosity, the fraud seemed out of character and came as a shock to her colleagues. *(Document: 5c73c9a927a4c)*
- A new IT system was brought in which would have made it harder to manipulate the records, therefore making it harder to commit fraud, but the manager used her influence to give her department exemption from using the new system. *(Document: 5c6ecc9f2513e)*
- The manager appeared kind, generous and supportive of people in her group. *(Document: 5c6e877f639e8)*
- The actions of the manager, who was described as kind and generous by one of her colleagues, was discovered after a teller at the bank questioned a cheque she had written for over $400,000. *(Document: 5c6e80c32e77a)*
- Colleagues perceived the manager as kind and generous, but there were some jokes and rumours about where they got their money from. *(Document: 5c6d857487398)*
- A middle manager, female, known as kind and generous/nice and understanding committed financial fraud. *(Document: 5c6d530129acc)*

Fig. 4. Topic 265

Topic 146

Closely related to: Actor Characteristics - Personality characteristics

- She probably wanted the money, but she was a nice and generous person with her employees, like buy food and drinks and help those who needed help. *(Document: 5d079eb114fb4)*
- All of her coworkers thought she was a nice person. *(Document: 5d056c1ca9952)*
- Everybody thought she was weird, but was nice. *(Document: 5d03fbe74cb6c)*
- When uncovered, everyone was surprised because the employee was always nice and generous towards everyone and could never imagined her stealing money for so long. *(Document: 5d03d561c08b4)*
- Staff thought she was a little quirky and sometimes joked about where she got her money, but they never did anything about it because she bought the occasional round at the bar and did other nice things for staff. *(Document: 5d03cbf7e42dc)*
- Co-workers were shocked as she seemed really nice. *(Document: 5d035b56d3d10)*
- Her colleagues were somewhat surprised by this as she always came across as a nice person to be around, although some did point at her being a bit flaky. *(Document: 5c6fee6c3665a)*
- Her colleagues had viewed her as being nice, caring and happy. *(Document: 5c6e9bcfbe44c)*
- The manager was seen to be a nice person, always willing to help and be there for her staff - even though some of them thought she was a 'bit flaky'. *(Document: 5c6d591491dac)*

Fig. 5. Topic 146

true (particularly in organic narratives) there is also a causal relationship which may not be directly temporally correlated.

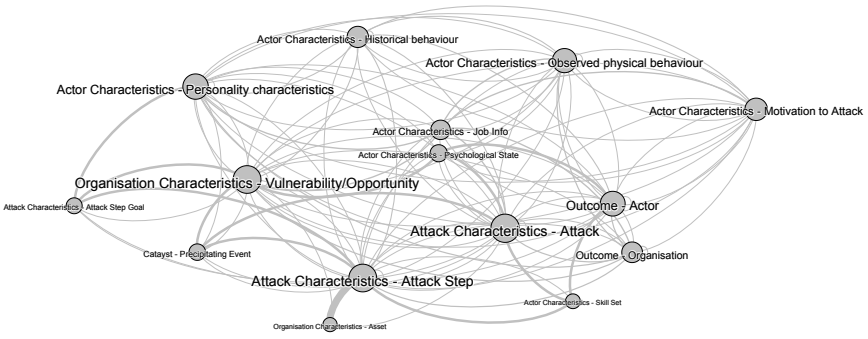

Fig. 6. The topics mapped with the naive approach

5 Discussion

The results presented demonstrate the model's ability to map sentences from organic narrative reports written by non-experts to an existing framework for modelling insider threat. This is shown by exploring various topics, their links to the insider threat model and the sentences within. However there are some areas for potential improvement, firstly some topics do not map well to the existing insider threat model, some topics link to multiple elements of an insider threat model and there are some mistakes made by the model.

In general the sentences match well with the models description of each element, following similar structures, further demonstrating the effectiveness of this technique. For example the element 'Attack Characteristics Attack' is defined in the initial report from which the case study was drawn [27] as 'Manipulating Company Records', although this is more general the majority of sentences within topics that link to this characteristic are related to the overall idea of writing fraudulent cheques as a manager in the tax office seen in topic 340 in Figure 7. This is also true for the element 'Organisation Characteristics - Vulnerability/Opportunity' which was described in [27] as 'Manual records, easily manipulated' and 'Inadequate security and processing (of records)', in Topic 132 in Figure 8 we see sentences regarding the auditing of records, a clear example of the inadequate security and sentences regarding the paper based system in use, once again an example of manual records. These two examples clearly demonstrate the unsupervised computational approach identifying similar characteristics that the authors of the initial case study expected to identify.

Topic 340

Closely related to: Attack Characteristics - Attack

- She was only caught because a bank teller questioned a cheque she wrote. *(Document: 5d07bded75b16)*
- There was a person who was a manager at the bank that some would describe as a nice woman who had been stealing millions from the bank overtime. *(Document: 5d03c707b6dfa)*
- She ended up getting caught when a bank teller marked a check as suspicious. *(Document: 5d039950a0ffa)*
- the bank manager was stealing checks she got away with it by telling everyone it was a family inheritance she was able to hide the transactions becasue the bank used a paper based system she was caught by a bank teller who thought a check looked suspicious the bank manager was caught and fined 60 million dollars *(Document: 5d028986a93c6)*
- A female bank manager was caught stealing money from the bank. *(Document: 5c7939498b4b6)*
- generally well-liked and popular female tax office manager had been stealing from taxes over 18 years by exploiting loopholes in paperwork systems and was very against an electronic system which she probably knew would make her theft harder to carry out and easier to detect, and was caught by a bank teller who noticed/questioned a suspicious cheque for $400,000; there were at least nine other accomplices; required to pay at least $45 million in restitution/taxes and other costs *(Document: 5c6fc0970bcac)*
- A bank teller helped catch her when they spotted a suspicious cheque for $400'000. *(Document: 5c6ecc9f2513e)*

Fig. 7. Topic 340

Topic 132

Closely related to: Organisation Characteristics - Vulnerability/Opportunity

- In the world new system to manage financial transaction, but we have worked old auditing and accounting paper systems, So did not provide & operating new system. *(Document: 5d03c9147ba66)*
- This old system did not have any controls for auditing or inspection. *(Document: 5c77c731bba1c)*
- However it was not straightforward to discover the extent of her fraudulent activities as there was no clear audit trail for her as she had been operating outside the computer system. *(Document: 5c73c9a927a4c)*
- Person was a female middle-manager at a company, operating with the assistance of nine others, managing to abuse the paper auditing (?) system to steal large amounts of money. *(Document: 5c6fc9bcf2c2e)*
- From a work perspective she was deemed experienced, important and knowledgeable enough to be involved in the creation of the new auditing system, but when it was implemented and she protested that it was not workable within her group, and despite the IT Group's insistence, did not have to work to it. *(Document: 5c6d6e434c6a8)*

Fig. 8. Topic 132

As discussed above some topics do not map well to the insider threat model, and therefore may not be associated with a characteristic within the insider threat framework. However this may also highlight a missing piece of a framework, emerging factors of an incident which have not yet been considered or as parts of other characteristics which have not been fully understood. An example from this case study is that the computational approach separated the 'outcome' theme from the framework into an outcome associated with the actor (the perpetrator of the attack) and an outcome associated with the organisation. This is is an interesting reflection with respect to our understanding of insider threat.

In addition to the issues with assigning topics to characteristics, another issue is the mislabelling of some sentences. For example, consider Topic 84 shown in Figure 9, the majority of sentences refer to the insider being caught, investigated, sent to court and asked to pay a fine, however, the final sentence 'The computer system was difficult to use and tax office staff found it an extra burden' is clearly an outlier. Although this is an issue, many of the mislabelled sentences are semantically different, this allows these sentences to be filtered out from the overall topic. To reduce the number of these sentences this we take the approach representing the sentences as a graph using co-reference resolution to join matching actors such as 'co-workers' to 'her office', and 'accomplices' to 'co-fraudsters'. The directed graph from these co-references would create highly connected graph referring to the intended characteristic, and a disjointed subgraph referring to the computer system and tax office.

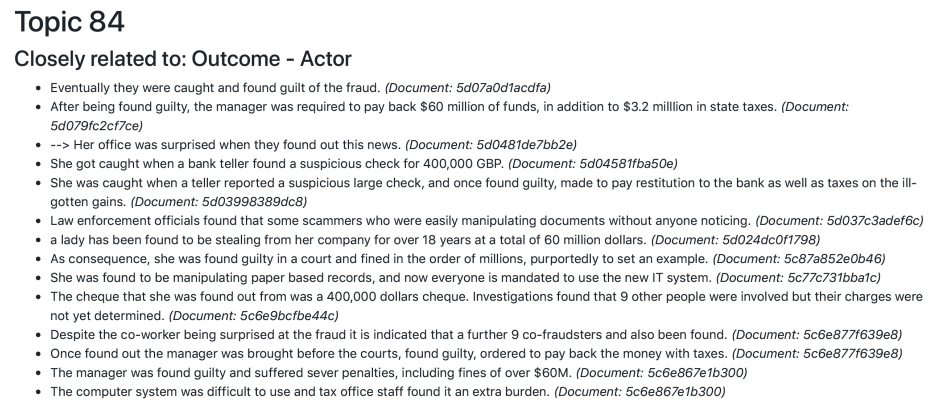

Fig. 9. Topic 84

It is clear that there are potential improvements that can be made, however the results still demonstrate the ability of topic models to computationally map a set of non-technical organic narratives to an existing technical security framework. Using topic modelling allows for additional advantages such as a model evolving as new reports are added, improving the model over time for a specific organisation.

The initial work in reconstructing a framework demonstrates that it is possible to link these topics together and to create a custom insider threat framework. Although further work needs to be done to explore causal links or temporal links, initial work shows that strongly linked characteristics already exist in the framework.

6 Conclusion

In this paper we have demonstrated an approach using Natural Language Processing (NLP) to computationally map organic narrative reports of insider threat attacks written in 'natural' language to an existing insider threat framework. This significantly reduces the barriers to gathering and generating actionable insight from a wide range of employees within an organisation. Reducing the cognitive load and the requirement for security knowledge we can improve the breadth of viewpoints of the incident and also reduce the effect of any confirmation bias in the model synthesis and hence improve the accuracy of a post-hoc model representation of the incident. In turn, this improved model representation improves the evidence used to generate an organisation's response to an incident with the ultimate aim of making organisations more secure.

By empowering the entire employee base to engage in an exercise, it is also possible to generate a more insightful study of an incident, it is also possible to hypothesis a study where an entire employee base write a short piece of prose of how they would compromise an organisation. These would form an interesting set of narratives that could be used to generate hypothetical models which represent the 'everyday' vulnerabilities that employees note as they go about their daily business.

This work forms a small part of a larger project to use NLP in understanding the threat from insider activity. The aim of which is to create a custom framework for each incident, which can merge, grow and evolve as the organisation experiences different attacks. With the ultimate goal of helping organisations develop appropriate and proportionate security decision to manage the risk from insider attack whilst empowering the entire employee-base to support the security of the organisation.

References

1. Agrafiotis, I., Nurse, J.R., Buckley, O., Legg, P., Creese, S., Goldsmith, M.: Identifying attack patterns for insider threat detection. Computer Fraud and Security **2015**(7), 9–17 (2015). https://doi.org/10.1016/S1361-3723(15)30066-X
2. Arun, R., Suresh, V., Madhavan, C.V., Murthy, M.N.: On finding the natural number of topics with latent dirichlet allocation: Some observations. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining. pp. 391–402. Springer (2010). https://doi.org/10/fndkt7
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research **3**(Jan), 993–1022 (2003)

4. Brown, C.R., Watkins, A., Greitzer, F.L.: Predicting Insider Threat Risks through Linguistic Analysis of Electronic Communication. 2013 46th Hawaii International Conference on System Sciences pp. 1849–1858 (2013). https://doi.org/10/gdrb3z
5. Butts, J.W., Mills, R.F., Baldwin, R.O.: Developing an insider threat model using functional decomposition. In: Gorodetsky, V., Kotenko, I., Skormin, V. (eds.) Computer Network Security. pp. 412–417. Springer Berlin Heidelberg, Berlin, Heidelberg (2005)
6. Cappelli, D., Moore, A., Trzeciak, R.: The CERT Guide to Insider Threats: How to Prevent, Detect, and Respond to Information Technology Crimes (Theft, Sabotage, Fraud). Addison-Wesley Professional (2012)
7. Coles-Kemp, L., Theoharidou, M.: Insider threat and information security management. In: Probst, C.W., Hunker, J., Gollmann, D., Bishop, M. (eds.) Insider Threats in Cyber Security. pp. 45–71. Springer US, Boston, MA (2010). https://doi.org/10.1007/978-1-4419-7133-3_3, https://doi.org/10.1007/978-1-4419-7133-3_3
8. Eberle, W., Graves, J., Holder, L.: Insider Threat Detection Using a Graph-Based Approach. Journal of Applied Security Research **6**(1), 32–81 (2010). https://doi.org/10.1080/19361610.2011.529413
9. Elmrabit, N., Yang, S.H., Yang, L.: Insider threats in information security categories and approaches. In: 2015 21st International Conference on Automation and Computing (ICAC). pp. 1–6 (2015). https://doi.org/10.1109/IConAC.2015.7313979
10. Forte, L.: Insider Threat Report 2019. Insider Threat Report 2019, Red Goat Cyber Security (2019)
11. Gavai, G., Sricharan, K., Gunning, D., Hanley, J., Singhal, M., Rolleston, R.: Supervised and unsupervised methods to detect insider threat from enterprise social and online activity data. Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA) **6**(4) (2015). https://doi.org/10.1145/2808783.2808784
12. Greitzer, Kangas, Noonan, Brown, Ferryman: Psychosocial Modeling of Insider Threat Risk Based on Behavioral and Word Use Analysis. e-Service Journal **9**(1), 106 (2013). https://doi.org/10/gdrb4d
13. Greitzer, F.L., Strozer, J., Cohen, S., Bergey, J., Cowley, J., Moore, A., Mundie, D.: Unintentional insider threat: Contributing factors, observables, and mitigation strategies. In: 2014 47th Hawaii International Conference on System Sciences. pp. 2025–2034 (Jan 2014). https://doi.org/10.1109/HICSS.2014.256
14. Greitzer, F.L., Hohimer, R.E.: Modeling human behavior to anticipate insider attacks. Journal of Strategic Security **4**(2), 25–48 (2011), http://www.jstor.org/stable/26463925
15. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proceedings of the National academy of Sciences **101**(suppl 1), 5228–5235 (2004)
16. Hirschberg, J., Manning, C.D.: Advances in natural language processing. Science **349**(6245), 261–266 (2015). https://doi.org/10/f7kfrk
17. Ho, S.M., Hancock, J.T., Booth, C., Burmester, M., Liu, X., Timmarajus, S.S.: Demystifying insider threat: Language-action cues in group dynamics. In: Proceedings of the Annual Hawaii International Conference on System Sciences. vol. 2016-March, pp. 2729–2738 (2016). https://doi.org/10.1109/HICSS.2016.343
18. Jacobi, C., van Atteveldt, W., Welbers, K.: Quantitative analysis of large amounts of journalistic texts using topic modelling. Digital Journalism **4**(1), 89–106 (Jan 2016). https://doi.org/10/f3s2sg

19. Johnston, A.C., Warkentin, M., McBride, M., Carter, L.: Dispositional and situational factors: influences on information security policy violations. European Journal of Information Systems **25**(3), 231–251 (2016). https://doi.org/10.1057/ejis.2015.15, https://doi.org/10.1057/ejis.2015.15
20. Kandias, M., Mylonas, A., Virvilis, N., Theoharidou, M., Gritzalis, D.: An insider threat prediction model. In: Katsikas, S., Lopez, J., Soriano, M. (eds.) Trust, Privacy and Security in Digital Business. pp. 26–37. Springer Berlin Heidelberg, Berlin, Heidelberg (2010)
21. Lo, R.T.W., He, B., Ounis, I.: Automatically building a stopword list for an information retrieval system. In: Journal on Digital Information Management: Special Issue on the 5th Dutch-Belgian Information Retrieval Workshop (DIR). vol. 5, pp. 17–24 (2005)
22. Maasberg, M., Warren, J., Beebe, N.L.: The dark side of the insider: Detecting the insider threat through examination of dark triad personality traits. In: 2015 48th Hawaii International Conference on System Sciences. pp. 3518–3526 (Jan 2015). https://doi.org/10.1109/HICSS.2015.423
23. Magklaras, G., Furnell, S.: Insider threat prediction tool: Evaluating the probability of it misuse. Computers & Security **21**(1), 62 – 73 (2001). https://doi.org/https://doi.org/10.1016/S0167-4048(02)00109-8, http://www.sciencedirect.com/science/article/pii/S0167404802001098
24. Manning, C.D., Manning, C.D., Schütze, H.: Foundations of Statistical Natural Language Processing. MIT press (1999)
25. Meng, F., Lou, F., Fu, Y., Tian, Z.: Deep learning based attribute classification insider threat detection for data security. In: 2018 IEEE Third International Conference on Data Science in Cyberspace (DSC). pp. 576–581 (June 2018). https://doi.org/10.1109/DSC.2018.00092
26. Nickerson, R.S.: Confirmation bias: A ubiquitous phenomenon in many guises. Review of General Psychology **2**(2), 175–220 (1998). https://doi.org/10.1037/1089-2680.2.2.175, https://doi.org/10.1037/1089-2680.2.2.175
27. Nurse, J.R.C., Buckley, O., Legg, P.A., Goldsmith, M., Creese, S., Wright, G.R.T., Whitty, M.: Understanding insider threat: A framework for characterising attacks. In: 2014 IEEE Security and Privacy Workshops. pp. 214–228 (May 2014). https://doi.org/10.1109/SPW.2014.38
28. Paxton-Fear, K., Hodges, D., Buckley, O.: Corpus expansion using topic modelling: Creating a library of insider threat attacks. Human-centric Computing and Information Systems (In Review)
29. Riedl, M., Biemann, C.: Topictiling: a text segmentation algorithm based on lda. In: Proceedings of ACL 2012 Student Research Workshop. pp. 37–42. Association for Computational Linguistics (2012)
30. Smeaton, A.F.: Using NLP or NLP resources for information retrieval tasks. Springer (1999)
31. Trilla, A.: Natural language processing techniques in text-to-speech synthesis and automatic speech recognition. Departament de Tecnologies Media pp. 1–5 (2009)
32. Verizon: 2019 data breach investigations report, https://enterprise.verizon.com/en-gb/resources/reports/dbir/